## 2.5 RNA structure

Like DNA, RNA is a hetero-polymer composed from nucleotides, each consisting of a sugar, a phosphate and a nucleic acid base. The sugar ribose in RNA has one more OH group, compared to deoxyribose in DNA. The four distinct bases in DNA are adenine (A), guanine (G), cytosine (C) and thymine (T), while in RNA uracil (U) takes the place of thymine. While the central role of DNA is storage of genetic information, RNA molecules carry a variety of roles from structural (as in the protein building machinery of ribosome) to information transfer (in messenger RNA). Concomitant with their diverse roles, the structure of RNA molecules is more complicated and they can assume a variety of shapes. An important distinction to DNA that enables such diversification is that RNA is a single stranded molecule. While two complimentary strands of DNA wrap around each other to form a stable and relatively rigid molecule, the single strand of RNA is more flexible. The molecule can fold upon itself bringing bases far apart along the backbone of the molecule close enough to form complimentary Watson-Crick pairs. While the primary structure refers to the sequence of bases along RNA, its secondary structure indicates the bases that come into contact to form complimentary bonds. The thus connected macromolecule then assumes particular shape(s) in three dimensions, known as its tertiary structure.

Given the sequence of RNA, can we predict its secondary structure? In principle one should list all possible pairings, compute their energies say by adding specified energies for the different Watson-Crick pairings), and select the lowest energy ones (or assign each folding an appropriate Boltzmann probability). For $N$ base pairs there a maximum of $N/2$ base pairings, which (ignoring constraints) can occur in $(N-1)(N-3)\cdots = (N-1)!!$ possible ways. Including the possibility that some bases are unpaired will increase the above number of states even further. However, for large $N$, $(N-1)!! \approx (N/e)^{N/2}$ which far exceeds the total number of possible arrangements of a polymer, which as we have seem grows at most as $g^N$. Thus a large fraction of pairings is excluded by steric constraints of foldability into a viable three dimensional structure. Although the number of foldable states still grows quite rapidly for large $N$, there are a number of algorithms that perform this computational task *for specific subsets of pairings* in polynomial time. A particularly convenient subset of pairings is that of planar graphs for which the RNA backbone, and all secondary connections can be drawn on a two dimensional plane, without any two lines crossing. Secondary connections that violate planarity lead to three dimensional structures containing elements called *pseudoknots* which are very rare (though not impossible) in actual RNAs. Thus limiting the search to this subset is not too severe a restriction.

The advantage of the planar subset of pairings is that it can be represented in multiple ways, and importantly enables finding the optimal configuration in polynomial time. One simple representation, indicated above, is obtained by stretching the RNA along a straight line and connecting the paired monomers by arches. Two arches are either disconnected, or one is entirely enclosed by the other– the arches will not intersect for planar graphs. Another representation is in terms of parentheses: Starting from one end of the RNA sequence, an open parenthesis is placed when the first nucleotide of pair is encountered, the parenthesis is closed when its partner is encountered. A planar diagram will then correspond to a gram-

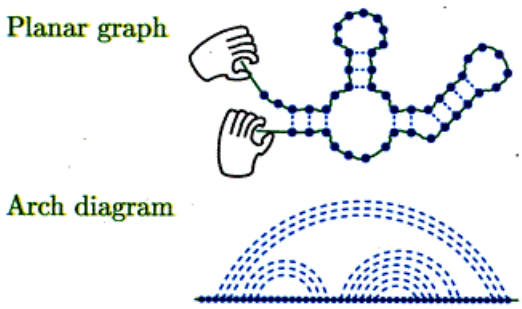


Figure 19: RNA secondary structure without pseudoknots represented as a planar graph or arch diagram.

matically correct string. The latter representation then yields a useful graphical prescription as a random walk: Moving along the sequence an up step indicates a parenthesis opened, a down step one that is closed. The planar diagram is now depicted as an island or mountain landscape with no segments where the height is negative.

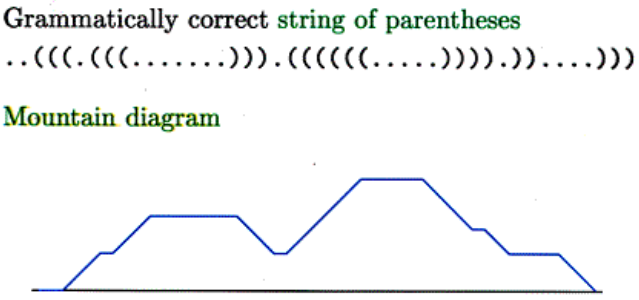


Figure 20: RNA secondary structure without pseudoknots represented with parenthesis, and as a mountain landscape.

Let us consider a simple model for secondary structures in which *all pairings* without pseudo-knots are allowed (i.e. without consideration of bending or steric constraints). For each configuration $C$, the energy is the sum over energies assigned to all bonded pairs, i.e. $E[C] = \sum_{<ij>\in C} \varepsilon_{ij}$, where $\varepsilon_{ij}$ is the energy of the bond between monomers $i$ and $j$; naturally the sum includes only the subset of indices paired in configuration $C$. The configuration of minimal $E$ can be obtained recursively as follows. Suppose we have found optimal configurations (and energies) for all sub-sequences of length $n$ and shorter. The optimal energy for a sub-sequence of length $(n+1)$, say spanning sites $i$ to $j = i+n+1$ is obtained by considering the following $(n+2)$ possibilities: $j$ (the last point) is unpaired in the optimal configuration, or $j$ is paired to a site $i \le k \le j-1$. In any one of the latter $(n+1)$ cases the arch between $j$ and $k$ creates two segments (from $i$ to $k-1$, and from $k+1$ to $j-1$) which are independent due to the planarity restriction. Since the optimal energies for the any of the latter two segments are known, the best energy is obtained by comparison

of all such possibilities as

$$E_{i,j} = \min\left[E_{i,j-1}, \varepsilon_{kj} + E_{i,k-1} + E_{k+1,j-1}\right] \quad \text{for} \quad i \le k \le j-1\,. \tag{2.97}$$

Starting from segments of length $n = 1$, where $E_{i,i+1} = \varepsilon_{i\ i+1}$, the above equation can be used to generate optimal energies for longer segments. The optimal configuration can then be obtained by tracing back. The number of operations required to find the optimal secondary structure of a sequence of length $N$ grows only as a $N^3$.

The above procedure is easily extended to finite temperatures where considerations of entropy may be relevant. We can then assign free energies to segments of the RNA, obtained from corresponding partition functions which may be computed recursively by appealing to Eq. (2.97) as

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} e^{-\beta\varepsilon_{kj}} Z_{i,k-1} Z_{k+1,j-1}\,. \tag{2.98}$$

### 2.5.1 Free energy of molten RNA

Using variants of Eq. (2.98), it is possible to follow how the secondary structure denatures as a function of temperature. Presumably at some temperature $T_m$ the native structure disappears in favor of a molten state resembling a branched polymer. The nature of the melting process should depend strongly on the RNA sequence and its native structure. In the next section we shall explore this melting for the simple case of an RNA hairpin. In its molten phase, RNA can explore a variety of structures reflecting the competition between energy gain of pairing and the resulting loss of entropy. To estimate the fraction of bound pairs in the molten phase, we can neglect variations in binding energy, setting $\varepsilon_{ij} = \varepsilon$ and a corresponding Boltzmann weight of $q \equiv e^{-\beta\varepsilon} \ge 1$. Once sequence variations are removed, the constrained partition function will depend only on the number of segment sites, i.e. $Z_{i,j} \equiv Z_m(|j-i|)$, and Eq. (2.98) simplifies to

$$Z_m(N+1) = Z_m(N) + q\sum_{k=1}^{N-1} Z_m(k-1) Z_m(N-k)\,, \quad \text{with} \quad Z_m(1) = 1\,. \tag{2.99}$$

It is possible to solve the above recursion relation by changing to an ensemble of variable length $N$. However, a more informative solution is obtained by considering the "mountain" representation of planar graphs. The correct weight for each graph is obtained by assigning a factor of 1 for each horizontal step, and $\sqrt{q}$ to a vertical step (up or down). Each configuration can then be regarded as a Markovian random walk with these weights, and the additional requirement that it never goes below the starting point. The constraint (for an island/mountain landscape, or correct formulation of parentheses) is thus equivalent to a barrier to the random walk at a position one step below the starting point. The problem of a random walk with a so-called absorbing barrier can be solved in several ways– a quite elegant solution is presented by Chandrasekhar in Rev. Mod. Phys. **15**, 1 (1943). Let us first ignore the constraint: The partition function for all paths of $N$ steps starting at the origin is

simply $(1+2\sqrt{q})^N$, accounting for all three possibilities in each step. Similarly, adding the uncorrelated fluctuations in each step leads to a variance of $\sigma^2 = N(2\sqrt{q})/(1+2\sqrt{q})$. In the limit of large $N$, and appealing to the central limit theorem, the net weight of the subset of walks ending at a height $h$ after $N$ steps is obtained as

$$W(N,h) = (1+2\sqrt{q})^N \exp\left[-\frac{(1+2\sqrt{q})h^2}{4\sqrt{q}N}\right] \times \sqrt{\frac{1+2\sqrt{q}}{4\pi\sqrt{q}N}}\,. \tag{2.100}$$

If we were to ask the question of what fraction of these random walks return to the origin ($h = 0$), we would obtain the expected result of $\Omega(N) \propto g^N/N^c$ with the 'loop closure' exponent of $c = 1/2$ for these one-dimensional random walks. Of course, for counting planar graphs we need the smaller subset of walks that return to the origin without ever passing to $h < 0$, and need to subtract all undesired walks from our sum.

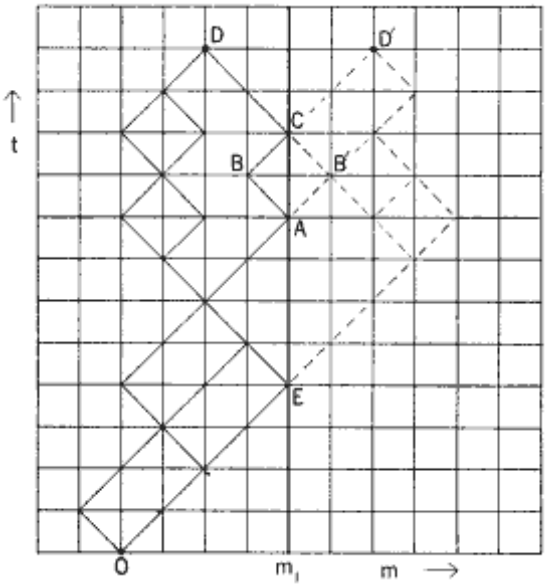


Figure 21: Copied from S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).

Chandrasekhar's solution to this problem is closely related to the method of images in electrostatics. The image of the starting point ($h = 0$) with respect to the forbidden state ($h = -1$) is located at $h = -2$. Consider all walks that start at this image point and end at $h = 0$. Each such walk $W^*$ must cross the 'mirror' plane at $h = -1$ at least once. We can construct a related ensemble of walks $W'$ which are the reflection of these walks in the mirror-plane (thus starting at the original point $h = 0$) up to the point of first intersecting the forbidden state at $h = -1$, and after which following the path of $W^*$. We note that the ensemble $W'$ consists of precisely the paths starting and ending at $h = 0$ which violate the non-crossing condition. As these paths are in one to one correspondence to $W^*$, we simply need to subtract them from the sum in Eq. (2.100) to get the correct number of non-crossing

paths. Since $W^*$ is the ensemble of walks with an end to end excursion of $h = 2$, we obtain

$$Z_m(N+1) = W(N,0) - W(N,2) = (1+2\sqrt{q})^N \left[1 - \exp\left(-\frac{(1+2\sqrt{q})}{\sqrt{q}N}\right)\right] \times \sqrt{\frac{1+2\sqrt{q}}{4\pi\sqrt{q}N}}\,. \tag{2.101}$$

The Gaussian approximation is only valid for large $N$, and we should similarly expand the difference in brackets above to get the asymptotic form

$$Z_m(N) \simeq A(q)\frac{g(q)^N}{N^c}\,, \quad \text{with } A(q) = \left(\frac{1+2\sqrt{q}}{64\pi^3\sqrt{q}}\right)^{3/2}\,,\ g(q) = (1+2\sqrt{q})\,,\ \text{and } c = \frac{3}{2}\,. \tag{2.102}$$

The most important consequence of the constraint is the change of the exponent $c$ from $1/2$ to $3/2$. The fraction of bound pairs is merely determined by the probability of going up or down at any step, and thus given by

$$\frac{\langle N_B\rangle}{N} = \frac{2\sqrt{q}}{1+2\sqrt{q}}\,, \tag{2.103}$$

which changes continuously from 1 at large $q$ (low temperatures) to $2/3$ as $q \to 1$ at high temperatures.

### 2.5.2 Melting of a hairpin

A particularly simple native RNA structure is a hairpin, where in its native configuration monomers $k$ and $2N-k+1$ are paired together. The denaturation of the hairpin involves the breaking of some of these bonds, potentially with formation of new bonds. For long hairpins the transition from the native form to the molten state can be described analytically using a so-called Gõ model[9]. In this model the native bonds are either assigned a stronger binding energy than non-native bonds. Possible configurations of the system are composed


Figure 22: The native state of a hairpin

of denatured segments that alternate with segments that maintain the original bonding.

A partition function is obtained by summing over all partially denatured configurations, and ignoring any interaction between the segments, takes the form

$$Z_n(N) = \sum_{l_1,l_2,l_3,\ldots}{}' R(l_1)Z_m(2l_2)R(l_3)Z_m(2l_4)\cdots\,, \tag{2.104}$$

[9]R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).

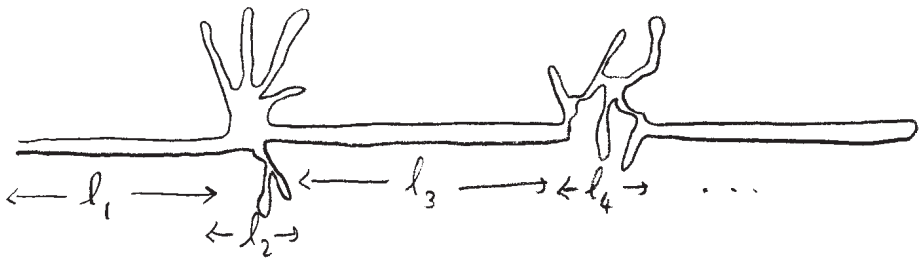


Figure 23: The molten state of a hairpin

with the constraint $l_1+l_2+l_3+\cdots = N$. The contribution of the molten segments come from Eq. (2.102). For the native segments, we should add the binding energies of the segments. To make the problem analytically tractable, we assign to each native bond an energy $\overline{\varepsilon} < \varepsilon$, and a corresponding Boltzmann weight $\overline{q} = e^{-\beta\overline{\varepsilon}} > q$.

With these simplifications, the problem becomes *identical* to the Poland–Scheraga model in Eq. (2.80) with $w = \overline{q}$, $g = 1 + 2\sqrt{q}$ and $c = 3/2$. It is thus possible to obtain a melting transition at a finite temperature at which the native fraction goes to zero linearly ($\beta = 1$) from Eq. (2.95)).